Partition models and cluster processes
With applications to classification

Peter McCullagh    Jie Yang

Department of Statistics
University of Chicago

TTI, Chicago, November 6 2006
Outline

1. Partition models and cluster processes
   - Fisher discriminant model
   - Logistic regression model
   - The Ewens partition process
   - The Gauss-Ewens cluster process

2. Part II: Classification using Cox processes
   - Cox process models
   - Permanent partition process
   - Superposition of permanent processes
   - Permanent partition models
   - Conditional relative intensity
Fisher Discriminant model

Description as a process (i.i.d.)

- **k** classes with frequencies \( \pi_1, \ldots, \pi_k \)
- Features \( X \) for class \( r \) are iid \( N(\mu_r, \Sigma) \) on \( \mathcal{X} = \mathbb{R}^q \)

Observations \( (Y_1, X_1), (Y_2, X_2), \ldots \) are i.i.d

\[
\text{pr}(Y_i = r) = \pi_r \text{ i.i.d.}
\]

Given \( Y \), features are iid \( X_j \sim N(\mu_{y_j}, \Sigma) \)

Density of \( (Y_i, X_i) \) at \( (y, x) \)

\[
\pi_y |\Sigma|^{-q/2} \exp \left( -\frac{(x - \mu_y)'\Sigma^{-1}(x - \mu_y)}{2} \right)
\]
Fisher Discriminant model
Application to classification

Estimation step:
\[ \hat{\pi}_r = \frac{n_r}{n}, \quad \hat{\mu}_r = \bar{x}_r, \quad \hat{\Sigma} = S; \]

Classification step:
for a new unit \( u' \) such that \( X(u') = x' \)
\[ p_r(Y(u') = r \mid \text{data}) \propto \pi_r \phi(x' - \mu_r; \Sigma) \]
(independent of data other than \( x' \))

Substitution step:
\[ \hat{p}_r(Y(u') = r \mid \text{data}) \propto n_r \phi(x' - \bar{x}_r; S) \]
(dependent on data through estimate)
Logistic regression model
Description as a process (indep but not i.d.)

Feature $x(u)$ regarded as a covariate for unit $u$
Class $Y(u)$ regarded as ‘response’

Linear logistic model:
Components $Y(u_1)$, $Y(u_2)$, … independent

$$\log \text{pr}(Y(u) = r) = \alpha_r + \beta_r x(u) + \text{const}$$
consistent with Fisher model $\beta_r = \Sigma^{-1} \mu_r$
Logistic regression model
Application to classification

Estimation step:
maximum likelihood gives $\hat{\alpha}, \hat{\beta}$

Classification step:

$$\Pr(Y(u') = r \mid \text{data}) = \frac{\exp(\alpha_r + \beta_r x(u'))}{\sum_c \exp(\alpha_c + \beta_c x(u'))}$$

(independent of data other than $x(u')$)

Substitution step: $\beta \mapsto \hat{\beta}$ (fitted = predicted)
Random partition of \([n] = \{1, \ldots, n\}\)

A set of disjoint non-empty subsets whose union is \([n]\)

- \(n = 2\): \(B = \{\{1, 2\}\} = 12\) or \(B = \{\{1\}, \{2\}\} = 1\vert 2\)
- \(n = 3\): \(B_3 = \{123, 1\vert 23, 2\vert 13, 3\vert 12, 1\vert 2\vert 3\}\)
- \(n = 4\): \(B_4 = \{1234, 123\vert 4\vert 4, 12\vert 34\vert 3, 12\vert 3\vert 4\vert 6, 1\vert 2\vert 3\vert 4\}\)

Bell numbers: \(\#B_3 = 5; \#B_4 = 15; \#B_5 = 52; \#B_6 = 203\)

\[ B = 12\vert 34\vert 5 \equiv \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \in B_5 \]
Permutation of elements: $B_{ij}^\sigma = B_{\sigma(i)\sigma(j)}$

$$B = 12|34|5 \equiv \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \in \mathcal{B}_5$$

Example 1: $\sigma = (1, 3)$: $B^\sigma = 32|14|5$
Example 2: $\sigma = (2, 4, 5)$: $B^\sigma = 15|23|4$

$B^\sigma(1, 5) = B(1, 2) = 1, \quad B^\sigma(1, 2) = B(1, 4) = 0$

Block sizes unaffected

Exchangeable product partition distribution (Hartigan 1990)

$$p_n(B) = D_n \prod_{b \in B} C(\#b)$$
Deletion of elements

Deletion (of 5) from $12\{34\}5$ gives $12\{34\}$ in $B_4$
Deletion (of 5) from $135\{2\}4$ gives $13\{2\}4$ in $B_4$

Deletion is a map $D : B_n \rightarrow B_{n-1}$
How does this affect distributions?

Example: $p_5$ is uniform on the 15 elements of $B_5$
Marginal distribution on $B_4$ after deletion is

\[
p_4(1234) = p_5(12345) + p_5(1234|5) = 2/15
\]
\[
p_4(123|4) = p_5(1235|4) + p_5(123|45) + p_5(123|4|5) = 3/15
\]
\[
p_4(12|3|4) = p_5(125|3|4) + \cdots + p_5(12|3|4|5) = 4/15
\]
Marginal distribution not uniform.
Ewens partition process

\[ p_n(B) = \frac{\lambda^{\#B} \Gamma(\lambda)}{\Gamma(n + \lambda)} \prod_{b \in B} \Gamma(\#b) \]

Exponential family of distns on \(B_n\)
- Canonical parameter \(\theta = \log \lambda\)
- Canonical statistic \(\#B\):
  \[ E(\#B) \sim \lambda \log(n) \]
- Cumulant function
  \[ \log \Gamma(n + e^\theta) - \log \Gamma(e^\theta) \]

Product partition model: \(C(\#b) = \lambda \times (\#b - 1)!\)
Finitely exchangeable for each \(n\)
\(p_n\) is marginal distn of \(p_{n+1}\)
distribution for $n = 4$

$$p_4(1234) = 6/(\lambda + 1)(\lambda + 2)(\lambda + 3),$$
$$p_4(123|4) = 2\lambda/(\lambda + 1)(\lambda + 2)(\lambda + 3),$$
$$p_4(12|34) = \lambda/(\lambda + 1)(\lambda + 2)(\lambda + 3)$$
$$p_4(12|3|4) = \lambda^2/(\lambda + 1)(\lambda + 2)(\lambda + 3),$$
$$p_4(1|2|3|4) = \lambda^3/(\lambda + 1)(\lambda + 2)(\lambda + 3)$$

$$p_3(123) = p_4(1234) + p_4(123|4) = 2/(\lambda + 1)(\lambda + 2)$$
$$p_3(12|3) = p_4(124|3) + p_4(12|34) + p_4(12|3|4)$$
$$p_3(1|2|3) = p_4(14|2|3) + p_4(1|24|3) + p_4(1|2|34) + p_4(1|2|3|4)$$

$p_n$ is marginal distribution of $p_{n+1}$

Infinitely exchangeable partition process
Ewens process: $B \equiv B_{\infty}$ is a random partition of natural numbers. Regard as a sequence of matrices $B_1, B_2, \ldots$ such that $B_n(i, j) = B_{\infty}(i, j)$ is leading sub-matrix of order $n$. 

$B_n \sim \text{Ewens}(n, \lambda)$

Conditional distribution of $B_{n+1}$ given $B_n$

Transition distribution $B_n \rightarrow B_{n+1}$

$$\Pr(u_{n+1} \mapsto b|B_n) = \begin{cases} 
\#b/(n + \lambda) & b \in B_n \\
\lambda/(n + \lambda) & b = \emptyset
\end{cases}$$

CRP interpretation due to Dubins and Pitman
No of occupied tables is approx Poisson($\lambda \log(n)$)
Gauss Ewens cluster processes

Pair \((B, X)\) such that \(X = (X_1, X_2, \ldots)\) (random sequence)
\(B = \{B_{ij}\}\) is a random partition of integers

Finite-dimensional distributions
\(p_n(B, x)\) is distn on \(\mathcal{B}_n \times \mathcal{X}^n\)
Invariant under permutation \(p_n(B^{\sigma}, x^{\sigma}) = p_n(B, x)\)
\(p_n\) is the marginal of \(p_{n+1}\)

\(B \sim \text{Ewens}_n(\lambda); \quad \text{given } B, X \sim N(0, \Sigma = I_n + \theta B)\)

\[
p_n(B, x) = \frac{\lambda^{\#B} \Gamma(\lambda)}{\Gamma(n + \lambda)} \prod_{b \in B} \Gamma(\#b) \times |2\pi\Sigma|^{-1/2} \exp(-x'\Sigma^{-1}x/2)
\]
Partition models and cluster processes
Part II: Classification using Cox processes
Fisher discriminant model
Logistic regression model
The Ewens partition process
The Gauss-Ewens cluster process
Potential applications of cluster process

- **Prediction**: given \((B, X)^{(n)}\) predict next value
  — Easily computed but not interesting or useful

- **Density estimation**: Given \(X_1, \ldots, X_n\) (No \(B\))
  predict next value — \(p_{n+1}(X_{n+1} \mid X_1, \ldots, X_n)\)

- **Cluster analysis**: \(p_n(B \mid X)\) is a distn on clusterings of \([n]\)
  \(p_n(B_{i,j} = 1 \mid X) = E(B_{i,j} \mid X)\) (triples too)
  \(p_n(\#B = r \mid X)\) (conditional distn of \(\#B\))
  — preliminary conclusions ...

- **Classification**: given \((B, X)^{(n)}\) and \(X(u_{n+1}) = x'\)
  compute \(p_{n+1}(u_{n+1} \mapsto b \mid \text{data})\)
  — Easily computed and fairly useful
Potential applications of cluster process

- **Prediction:** given \((B, X)^{(n)}\) predict next value
  — Easily computed but not interesting or useful

- **Density estimation:** Given \(X_1, \ldots, X_n\) (No \(B\)) predict next value — \(p_{n+1}(X_{n+1} \mid X_1, \ldots, X_n)\)

- **Cluster analysis:** \(p_n(B \mid X)\) is a distn on clusterings of \([n]\)
  \(p_n(B_{i,j} = 1 \mid X) = E(B_{i,j} \mid X)\) (triples too)
  \(p_n(\#B = r \mid X)\) (conditional distn of \(\#B\))
  — preliminary conclusions ...

- **Classification:** given \((B, X)^{(n)}\) and \(X(u_{n+1}) = x'\) compute \(p_{n+1}(u_{n+1} \mapsto b \mid \text{data})\)
  — Easily computed and fairly useful
Potential applications of cluster process

- **Prediction:** given \((B, X)^{(n)}\) predict next value
  — Easily computed but not interesting or useful

- **Density estimation:** Given \(X_1, \ldots, X_n\) (No \(B\))
  predict next value — \(p_{n+1}(X_{n+1} \mid X_1, \ldots, X_n)\)

- **Cluster analysis:** \(p_n(B \mid X)\) is a distn on clusterings of \([n]\)
  \[p_n(B_{i,j} = 1 \mid X) = E(B_{i,j} \mid X)\] (triples too)
  \[p_n(\#B = r \mid X)\] (conditional distn of \(\#B\))
  — preliminary conclusions ...

- **Classification:** given \((B, X)^{(n)}\) and \(X(u_{n+1}) = x'\)
  compute \(p_{n+1}(u_{n+1} \mapsto b \mid \text{data})\)
  — Easily computed and fairly useful
Potential applications of cluster process

- Prediction: given $(B, X)^{(n)}$ predict next value
  — Easily computed but not interesting or useful

- Density estimation: Given $X_1, \ldots, X_n$ (No $B$)
  predict next value — $p_{n+1}(X_{n+1} \mid X_1, \ldots, X_n)$

- Cluster analysis: $p_n(B \mid X)$ is a distn on clusterings of $[n]
  p_n(B_{i,j} = 1 \mid X) = E(B_{i,j} \mid X)$ (triples too)
  $p_n(\#B = r \mid X)$ (conditional distn of $\#B$)
  — preliminary conclusions ...

- Classification: given $(B, X)^{(n)}$ and $X(u_{n+1}) = x'$
  compute $p_{n+1}(u_{n+1} \mapsto b | \text{data})$
  — Easily computed and fairly useful
Given \((B, x)\) on a sample of \(n\) units and a new unit \(u'\) with \(x(u') = x'\).
Classify the new unit (conditional distribution)

Gauss-Ewens conditional distribution

\[
\Pr(u' \mapsto b \mid \ldots) \propto \begin{cases} 
(#b) \phi_{n+1}(x'; I_{n+1} + \theta B_b) & b \in B \\
\lambda \phi_{n+1}(x'; I_{n+1} + \theta B_{\emptyset}) & b = \emptyset
\end{cases}
\]

\[x' = (x, x'); B_b = \{B, u' \mapsto b\}\]
Application to classification (contd)

Simplification of conditional distribution

\[
\text{pr}(u' \mapsto b | \ldots) \propto \begin{cases} 
(\#b) \phi_1(x(u') - \tilde{\mu}_b; \tilde{\sigma}_b^2) & b \in B \\
\lambda \phi_1(x(u'); 1 + \theta) & b = \emptyset
\end{cases}
\]

\[
\tilde{\mu}_b = \bar{x}_b n_b \theta / (1 + n_b \theta), \quad \tilde{\sigma}_b^2 = 1 + \theta / (1 + n_b \theta)
\]

Typical values \( \theta \geq 5 \) and \( n_b \geq 5 \)
so \( \tilde{\mu}_b / \bar{x}_b = n_b \theta / (1 + n_b \theta) \geq 0.96 \)

(similar to Fisher discriminant model, but with shrinkage)
Cox process models

Construction:
\( \lambda(x) \) is a random intensity function on \( \mathcal{X} = \mathbb{R} \)
\( \mathbf{X} \) is a Poisson process with intensity \( \lambda \).

Product density:
\[
\begin{align*}
\lambda(x) &= E(\lambda(x)) \text{ first-order p.d.} \\
m(x_1, x_2) &= E(\lambda(x_1)\lambda(x_2)) \text{ second order p.d.} \\
m(x) &= E(\lambda(x_1) \cdot \cdot \cdot \lambda(x_n)) \text{ is } n\text{th order p.d.}
\end{align*}
\]

\( \mathbf{X}(A) = \#(\mathbf{X} \cap A) = \text{no of events in } A \)
\( m(x) \, dx: \text{expected no of ordered } n\text{-tuples of events in } d\mathbf{x} \)
Marked Cox process

\[ \lambda_1(x), \ldots, \lambda_k(x) \text{ indep random intensity functions} \]

\[ X^{(r)} \text{ Cox process with intensity } \lambda_r(x) \]

\[ X = \bigcup X^{(r)} \text{ superposition process} \]

Marked process is labelled partition \( X^{(1)}, \ldots, X^{(k)} \) of \( X \)

\[ \Pr(X^{(r)} \text{ contains event in } dx^{(r)}) = m_r(dx^{(r)}) + o(dx^{(r)}) \]

\[ \Pr(X \text{ contains event in } dx) = m(dx) + o(dx) \]

\[ \Pr(y \mid x \subset X) = \frac{\prod m_r(x^{(r)})}{m(x)} \]

label \( y : x \rightarrow C \) and \( x^{(r)} = y^{-1}(r) \)
Boson and related processes

$Z(x)$ is a GP $N(0, K)$ on $\mathcal{X}$ (complex-valued for simplicity!)

$\lambda(x) = |Z(x)|^2$ is intensity at $x$ for Cox process

Product density: $m(x) = E(\lambda(x_1) \cdots \lambda(x_n)) = \text{per}[K](x)$

Matrix $[K](x) = \{K(x_i, x_j)\}$

Permanent polynomial of matrix $A$:

$$\text{per}_\alpha(A) = \sum_{\sigma} \alpha^{\#\sigma} A_{1,\sigma(1)} \cdots A_{n,\sigma(n)}$$

Convolution or superposition:

$$\sum_{w \subset x} \text{per}_\alpha[K](w) \text{per}_{\alpha'}[K](\bar{w}) = \text{per}_{\alpha+\alpha'}[K](x)$$

Macchi 1975; Shirai and Takahashi 2003; McCullagh and Møller 2006
Partition models and cluster processes
Part II: Classification using Cox processes

Intensity functions and Cox processes

Cox process models
Permanent partition process
Superposition of permanent processes
Permanent partition models
Conditional relative intensity
Intensity functions and Cox processes
Superposition processes

Partition models and cluster processes
Part II: Classification using Cox processes

Superposition of permanent processes
Permanent partition models
Conditional relative intensity

Peter McCullagh
Given that $\mathbf{x} \subset \mathbf{X}$ occurs in the superposition
Conditional distribution of labels $\mathbf{y} : \mathbf{x} \rightarrow \mathcal{C}$

$$p_n(\mathbf{y} \mid \mathbf{x}) = \frac{\text{per}_{\alpha_1}[K](\mathbf{x}^{(1)}) \cdots \text{per}_{\alpha_k}[K](\mathbf{x}^{(k)})}{\text{per}_{\alpha}[K](\mathbf{x})}$$

is a prob distribution on labelled partitions of $\mathbf{x}$. reduces to Dirichlet multinomial if $K(x, x') \equiv 1$

Classification distribution for new event at $x'$

$$p_{n+1}(\mathbf{y}(x') = r \mid \text{data}) \propto \text{per}_{\alpha_r}[K](\mathbf{x}^{(r)}, x') / \text{per}_{\alpha_r}[K](\mathbf{x}^{(r)})$$
Conditional relative intensity \( \text{per}_\alpha[K](x, x') / \text{per}_\alpha[K](x) \)

\[ K(x, x') = \exp(-(x - x')^2) \text{ on } (0, 2\pi). \]

Solid curve is for \( \alpha = 1 \), dashed curve for \( \alpha = 0.25 \), and the dotted curve for \( \alpha = 0 \).
Illustration 2: Relative intensity functions $p_{\text{per}}[K](\mathbf{x}, \mathbf{x}') / p_{\text{per}}[K](\mathbf{x})$

One-dimensional feature space: 2 labelled classes:
A two-class example is generated by a $3 \times 3$ chequerboard pattern. The permanent model with $\alpha_1 = \alpha_2 = \alpha$ and $K(x, x') = \exp(-\|x - x'\|^2/\tau^2)$ is used.
Chequerboard Pattern: Classification